# Automatic error detection in Russian learner language


| Elena Klyachko | Timofey Arkhangelskiy |
|---|---|
| National Research University Higher School of Economics | National Research University Higher School of Economics |
| elenaklyachko@gmail.com | tarkhangelskiy@hse.ru |

| Olesya Kisselev | Ekaterina Rakhilina |
|---|---|
| Portland State University | National Research University Higher School of Economics |
| kisselev@pdx.edu | erakhilina@hse.ru |


## 1 Introduction

Learner corpora, also known as interlanguage (IL) or second language (L2) corpora, have become increasingly popular resources in language research in the past decade. Learner corpora provide large volume of rich data for theoretical and applied language studies. Just as native (or L1) corpora, learner corpora are particularly useful for research when they are tagged; and learner corpora often contain tags that are more intricate than those found in L1 corpora. Metalinguistic tags, for instance, often contain information relevant both to the author of the text (language background, level, etc.) and the task (genre, format, time restriction, etc.). In regards to grammatical annotation, in addition to the usual lemmatisation and morphosyntactic tagging, L2 corpora may contain error-tags that provide information on deviant language use.

Error-tagging is known to be a resource-consuming and technologically-challenging task, more so for highly inflectional languages such as Russian, with its rich inflection, derivation, agreement, and robust syntax. Yet, the ability to conduct corpus searches by the types of errors may afford better insight into the processes of language acquisition; this goal warrants the necessity of error-tagging learner corpora. The project described in the present paper attempts to develop an automatic error-tagging protocol for Russian learner language. The annotation procedures are currently tested on the materials of the Russian Learner Corpus of Academic Writing (RULEC).

## 2 A learner corpus of Russian

RULEC (Alsufieva et al. 2012), a longitudinal developmental corpus of student writing, is a project initiated at the Russian Flagship Program at Portland State University (USA). It comprises academic papers written by advanced students of Russian as a Foreign Language (RFL) and students of Russian as a Heritage Language (RHL). All texts in the corpus were originally produced as regular class assignments; the current size of the corpus is 350,000 words. All texts contain a detailed meta-linguistic tag, providing information on the text author (with such variables as gender, L1, level, etc.) and the task (topic, genre, time restriction and the like). The corpus in its current state (with automatic grammatical markup, but without error annotation and error corrections) is available online.[1]

RULEC is also currently being lemmatised and annotated for grammatical information, – a challenging task in itself, given the high level of spelling inaccuracies (i.e. orthographic errors). Since automatic morphological tagging of learner language presumes dealing with errors, a fuller error-tagging was deemed a logical progression.

## 3 Automatic error-tagging of Russian learner language

The pilot study was conducted on a smaller portion of learner texts (apprx. 30,000 words). The first step in the process of error-tagging was error categorisation. The set of categories emerged as a result of initial manual annotation and currently includes orthographic, ortho-morphological, morpho-syntactic, syntactic and lexical errors. Manual error annotation is extremely time-consuming; in order to accommodate the growing volume of projects involving Russian learner and heritage corpora, a development of an automatic error-tagging system is a timely and necessary task.

An approach that this team takes to (partially) solve the difficulty inherent in automatic detection of non-orthographic errors in Russian (such as adjective and noun agreement) is essentially based on comparison of lists of bi- and tri-grams found in the learner corpus to lists of n-grams found L1 corpora. The actual process consists of a number of steps:

1. The learner texts are first checked by a Russian-language spellchecker, which marks the incorrect forms and corrects some of them. The statistics on the most frequent deviations is collected (such as the absence of a soft sign marking a palatalised consonant in *bol'shoi*). This statistics, if collected on the large volume of texts, can inform (and even predict) the frequency of the types of errors, that characterise texts produced by different groups learners (different levels, L1s, etc.).

1 http://web-corpora.net/RussianLearnerCorpus/search/

2. The bi-grams obtained from the texts from RULEC are then compared to the list of bi-grams found in the native corpora (such as Russian National Corpus). Such comparison allowed us to identify particular patterns, which in turn help better identify errors in particular types of constructions. Examples of such constructions are sequences "adjective+noun" (but not the other way around), in which the adjective almost always agrees with the noun, and constructions "preposition+noun," in which the preposition almost always governs the noun. Observations like these help define a set of rules which are then used to detect and correct such errors. With the help of morphological analyser, we check gender, number and case agreement in continuous sequences "adjective+noun," as well as prepositional government in sequences including "preposition+noun." Constructions that show errors in agreement and/or government are marked appropriately.

Nevertheless, there are notable exceptions to such rules. For example, the adjective is not in agreement with the noun in bi-grams like *neponyatnaya*.SG.NOM *chitatelyam*.PL.DAT 'obscure to the readers'. Thankfully, such exceptions will show up in the lists of frequent bi-grams, so in order to deal with them correctly we need to check the bi-gram supposed to be erroneous against the list of frequent authentic bi-grams.

3. In the next step, texts are checked against the lists of bi- and tri-grams obtained from from Google n-grams and the Russian National Corpus. Those text sequences that do not match any bi- and tri-grams in L1 texts or match only n-grams of very low frequency, are marked as (potentially) erroneous. Both bi-gram and tri-gram lookup are essential for error detection. Only tri-gram search would result in too many false positives, since far too many correct tri-grams will never appear on the list. On the other hand, using only bi-gram search is often not enough. For instance, three continuous words that happen to pair up as two bi-grams but never as a tri-gram in native data, should also be marked as erroneous. This process can aid in identification of non-standard Russian construction such as **postupal v institute* 'enrolled in the university' where a bi-gram *postupal v* and a bi-gram *v institute* are both highly-frequent but the exact tri-gram is impossible.

A word sequence is more likely to be an error if the list of authentic bi- and tri-grams contains a highly-frequent similar element (similar yet differing in either one or two graphic symbols, or one grammatical category of one of the words.) In this case, the errors may be corrected by using the most frequent "corrected" variant pulled from the list of corrected alternatives.

Although this approach proved to be productive on a pilot set of learner data, it is not always perfect: if, for instance, an incorrect word or sequence of words has a big Levenshtein distance to the respective correct form, the automated search for the probable correct form may not be successful. Allowing for the longer distance does not solve this problem, since it will scoop up too many variations and will slow down the process. This problem most often arises with forms that contain two or more typical errors (e.g. **bolshi* instead of the required *bol'she*). In this case, we employ the statistics on mis-spellings collected in Step 1. Instead of searching for all possible forms using Levenshtein metric, we only consider forms with typical (i.e. highly-frequent) errors.

## 4 Problems and limitations

So far, we see n-grams being successfully applied to identification of errors in, among others, noun-adjective agreement and prepositional government. Nevertheless, the method we employ poses certain limitations on which errors can be detected. The results are less successful in discontiguous structures (e. g. where the dependent word is one or two positions removed from the governing word). Using lists of n-grams collected from the authentic monolingual corpora of mostly literary texts from the 20th century may also be problematic, if, for instance, a particular bi- or tri-gram from a learner corpus happens to contain a more contemporary lexeme. Word order may also impact the content on an n-gram list, with learner texts containing examples of n-grams that are grammatical and perfectly correct in oral discourse, but very rare in the written natives data.

## 5 Conclusion

Error annotation – even (semi)automatic – is labor- and resource-intensive task; yet, the potential benefit of such work is significant. The possibility and the scope of studies done of error-annotated learner corpora may yield results and conclusions unavailable to language scholars and practitioners otherwise.

The use of n-grams to perfect corpus annotation process has become a widespread practice (Leacock et al. 2010). N-grams are used to deal with polysemy, errors of various nature, and more (Inkpen, Islam, 2010). This methodology can be applied to many different languages, including Russian and other less-frequently researched languages (unlike methods that use such platforms

as WordNet which are not available for most languages (Budanitsky and Hirst 2006)). Still, the most widely used procedures involving n-grams have been developed for languages other than Russian (or, for that matter, other Slavic languages), and cannot always be directly applied to these highly-inflectional languages. Hence, in order to use n-grams for error-detection in Russian, one must modify the procedures that may successfully work for English.

The proposed method for automatic error detection has been pilot-tested (without tri-grams) on a smaller subset of RULEC and proved its worth for Russian data. The preliminary results show that those of the most frequent non-orthographic errors which are best detected with this method include errors in prepositional and verbal government (80% detection rate), agreement in "noun+adjective" couples (90%), and errors of lexical choice (71%). Additionally, the results of n-grams comparisons allow the researchers to formulate explicit detection rules (for example, obligatory agreement in "adjective+noun" and "preposition+noun" sequences); these rules help streamline the automatic process of error detection in Step 2.

We see the potential in further developing the described approach to automatic error-detection in learner Russian. At the same time, our work is in its beginning stages; we expect to obtain constructive and plentiful feedback on the work presented in this paper, and to perfect our method as we learn from practice and colleagues.

## References


Alsufieva, A., Kisselev, O., and Freels, S. 2012. "Results 2012: Using Flagship Data to Develop a Russian Learner Corpus of Academic Writing". Russian Language Journal 62: 79-105

Budanitsky, A. and Hirst, G. 2006. "Evaluating WordNet-based measures of semantic distance". Computational Linguistics 32 (1): 13-47

Granger, S. 1998. Learner English on Computer. London: Addison Wesley Longman.

Inkpen, D. and Islam, A. 2010. "Unsupervised Approaches to Text Correction using Google n-grams for English and Romanian". In D. Tufis and C. Forascu (eds.) Multilinguality and Interoperability in Language Processing with Emphasis on Romanian: 270-285. Bucharest: Romanian Academy Publishing House

Leacock, C., Chodorow, M., Gamon, M. and Tetreault, J. 2010. Automated Grammatical Error Detection for Language Learners. Morgan & Claypool.